Chinese Journal of Electronics
Vol.29, No.6, Nov. 2020

CSELM-QE: A Composite Semi-supervised Extreme Learning Machine with Unlabeled RSS Quality Estimation for Radio Map Construction

ZHAO Jianli[1], WANG Wei[1], SUN Qiuxia[2], HUO Huan[3], SUN Guoqiang[1], GAO Xiang[1] and ZHU Chendi[1]

(1. *College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266510, China*)

(2. *College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao 266590, China*)

(3. *Faculty of Engineering and Information Technology, University of Technology Sydney, Australia*)

Abstract — Wireless local area network (WLAN) fingerprint-based localization has become the most attractive and popular approach for indoor localization. However, the primary concern for its practical implementation is the laborious manual effort of calibrating sufficient location-labeled fingerprints. The Semi-supervised extreme learning machine (SELM) performs well in reducing calibration effort. Traditional SELM methods only use Received signal strength (RSS) information to construct the neighbor graph and ignores location information, which helps recognizing prior information for manifold alignments. We propose Composite SELM (CSELM) method by using both RSS signals and location information to construct composite graph. Besides, the issue of unlabeled RSS data quality has not been solved. We propose a novel approach called Composite semi-supervised extreme learning machine with unlabeled RSS Quality estimation (CSELM-QE) that takes into account the quality of unlabeled RSS data and combines the composite neighbor graph, which considers location information in the semi-supervised extreme learning machine. Experimental results show that the CSELM-QE could construct a precise localization model, reduce the calibration effort for radio map construction and improve localization accuracy. Our quality estimation method can be applied to other methods that need to retain high quality unlabeled Received signal strength data to improve model accuracy.

Key words — Wireless local area network (WLAN), WiFi fingerprints, Radio map construction, Semi-supervised extreme learning machine (SELM), Received signal strength (RSS) data quality estimation, Location based services.

I. Introduction

The Wireless fidelity (WiFi) fingerprint-based localization has become a popular approach due to its wide deployment and availability of WiFi infrastructure. Fingerprint-based indoor localization is generally divided into offline phase and online phase. The radio map preprocessing is involved in the offline phase as an indispensable part in Ref.[1]. The offline phase constructs a radio map by recording the signal strength from different Access points (APs) in range and storing this information in a database along with the known coordinates of the mobile device. This phase involves many calibration effort to collect the Received signal strength (RSS) measures from all available access points at each known location. These known locations are defined as Reference points (RPs). The online phase estimates user location by comparing the current real-time RSS vector at an unknown location to those stored in the radio map and returns the closest match.

Localization accuracy is influenced by the collection density of radio map[2–4]. The heavy initial offline phase needs to record the signal strength from large number of RPs to construct an accurate radio map. Compared to labeled data, the collection of unlabeled RSS data without locations is relatively easy. Therefore, semi-supervised learning can solve the problem by using large amount of unlabeled data together with limited labeled data to reduce human effort and offer higher accuracy. In order to reduce the calibration effort of constructing

Manuscript Received Sept. 17, 2019; Accepted Mar. 12, 2020. This paper is supported by the Key R&D Plan of Shandong Province (No.2018GGX101045), the National Key R&D Plan (No.2018YFC0831002, No.2017YFC0804406), the Key Project of Industrial Transformation and Upgrading (Made in China 2025) (No.TC170A5SW), Humanity and Social Science Fund of the Ministry of Education (No.18YJAZH136, No.17YJCZH262), and the National Natural Science Foundation of China (No.61433012, No.U1435215).

DOI:10.1049/cje.2020.09.002

the radio map and ensure high localization accuracy, we propose the CSELM-QE (Composite semi-supervised extreme learning machine with unlabeled RSS quality estimation). The main contributions of this paper are summarized as follows. First, we combine the semi-supervised learning approach SELM (Semi-supervised extreme learning machine) with the composite neighbor graph that preserves the neighbor relations between location space and RSS space to improve localization accuracy and reduce data calibration effort at the same time[5]. Second, we consider the quality of unlabeled RSS data, which may bring noise to localization and design a quality estimation method to select high quality RSS data for SELM. Our Quality Estimation method can also be applied to other unlabeled RSS data processing algorithms, which need to retain high quality data to improve model accuracy. Finally, we conduct sufficient experiments in a real-world environment to prove that our method can guarantee high localization accuracy especially in the sparse environment compared with other methods.

The remainder of the paper is organized as follows: Section II reviews related works. Section III details the basic algorithm used in the current study and introduces the CSELM-QE approach. Section IV presents how to select the parameters for indoor localization and the comparative experimental results. Finally, Section V draws the conclusion and future work.

II. Related Works

Many researchers have worked to reduce manual effort in fingerprint-based localization. The common method is based on machine learning, which is applied in many fields, such as location-based recommendation services[6]. Traditional methods are triangulation interpolation, linear interpolation and Kriging interpolation[7]. However, when the fingerprints are sparse, these methods can not capture the random properties such as signal propagation characteristic, which leads to low localization accuracy. Other regression-based methods, such as Gaussian process regression and support vector regression[8,9] estimate the expected RSS at non-site-surveyed locations to reduce the site survey effort.

Compared with the labeled RSS data, the unlabeled RSS data without locations are easily to collect. Therefore, another research direction focuses on semi-supervised manifold learning that utilizes a large number of unlabeled and few of labeled RSS data to realize localization without increasing calibration effort[10,11]. Liu *et al.*[12] proposed the SELM, which uses graph Laplacian regularization to (Extreme learning machine) ELM to train a precise model for localization. Zhou *et al.*[5] used graph-based semi-supervised manifold alignments for localization, which preserves the neighbor relations of data both in signal and physical space. These methods all import unlabeled data for manifold learning and made a progress in reducing manual effort. However, a large number of unlabeled data also bring noise in semi-supervised manifold learning, which reduces the localization accuracy.

Other researchers have used user traces with unlabeled data and limited labeled data to improve localization accuracy[13,14]. In addition, researchers have tried crowdsourcing methods to collect training data at low cost. The crowdsourcing methods collect RSS measurements and use un supervised learning to estimate locations[15,16]. Apart from using RSS measurements, many of them use time stamps and inertial sensors (*e.g.*, accelerometer, compass, and gyroscope) embedded in smart phones to collect the sensing data on step number, direction, and angular rate. Rai *et al.*[17], Yang *et al.* and Wu *et al.*[18,19] estimated the relative locations of RSS measurements using inertial sensors and obtaining the absolute locations by matching these users traces with indoor map. Although these methods can reduce the calibration effort to a certain extent, they need to obtain the locations by relying on additional sensors. In addition, they have problems with inaccurate sensors, quality control and fusion of sensing data and device diversity which could affect localization accuracy. Several researchers focuses on data quality estimation for crowdsensing. Liu *et al.*[20] trained a context-aware classifier using historical data to estimate the quality of sensing data based on the Gaussian mixture model (GMM). Yang *et al.*[21] integrated quality estimation and monetary incentive, and proposed a truth estimation and surplus sharing method for crowdsensing environment.

Based on above works, we use the semi-supervised learning method SELM to learn the semi-supervised manifold and reserve location information to construct neighbor graphs to train the learning model. In addition, we take into account the quality of all the unlabeled RSS data and propose the method CSELM-QE with data quality estimation to improve localization accuracy.

III. CSELM-QE

This section first describes the problem we aim to solve, and then introduces the basic theory of graph based SELM algorithm. Secondly, we introduce the neighbor graph incorporating the physical location relations of labeled received signal strength data to improve Semi-supervised extreme learning machine. Finally, we introduce the method to evaluate the unlabeled received signal strength data to filter the noised ones.

Our proposed algorithm can be summarized as shown in Fig.1:

1. Fundamental definition of proposed algorithm

A localization process can be described as a regression problem $Y = f(x)$, the input variable is the vector of received signal strength and Y is the physical coordinate. And the problem is to train the regression model f. For indoor localization, high accuracy needs adequate calibration. Therefore, accurate localization with low calibration effort is an important issue. Semi-supervised machine learning methods have been applied by importing large number of unlabeled received signal strength data, which are easily collected. In addition, the ELM is proved to be good at achieving regression accuracy. To reduce the calibration workload, we use the semi-supervised extreme machine algorithm combined with composite graph construction and unlabeled RSS data quality estimation to reduce manual effort and ensure localization accuracy.

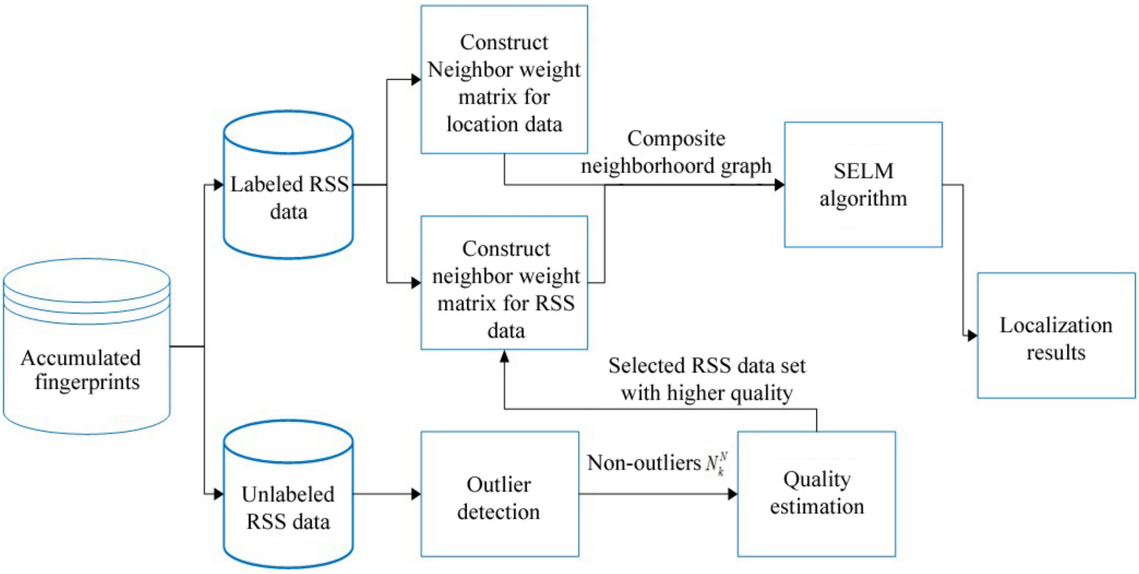

Construct Neighbor weight matrix for location data
Labeled RSS data
Composite neighborhoord graph
SELM algorithm
Construct neighbor weight matrix for RSS data
Accumulated fingerprints
Localization results
Selected RSS data set with higher quality
Unlabeled RSS data
Outlier detection
Non-outliers N_k^N
Quality estimation

Fig. 1. The architecture of proposed algorithm

Suppose there are N reference points selected and M WiFi access points visible in indoor areas. In the offline phase, we calibrate l RPs in the target environment and collect the corresponding RSS data as $x_i = \{rss_{i1}, rss_{i2}, \ldots, rss_{iM}\}(1 \leq i \leq l)$, the physical location as $l_i = (p_i, q_i)^{\mathrm{T}}$. We refer to these data as labeled RSS data. The users can hold mobile phones and walk in the wireless environment to collect u RSS data with unknown locations as the unlabeled RSS data. By setting l and u as the number of labeled data and unlabeled data respectively, the input signal space can be represented as $X = \{x_1, x_2, \ldots, x_{l+u}\}$, and the output matrix $T = \{l_1, l_2, \ldots, l_{l+u}\} \in R^{(l+u)\times 2}$. These data definitions are used in the following methods in our algorithm.

2. Basic methods

1) Semi-supervised extreme learning machine

Huang *et al.*[22] proposed ELM as a novel learning method. It is a Single hidden layer feedforward neural network (SLFN). The output of a neural network with L hidden nodes can be represented as follows:

$$f_L(x_i) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x_i), \ i = 1, 2, \ldots, N \quad (1)$$
$$\text{s.t. } a_i \in R^n, b_i \in R, \beta_i \in R^m$$

where n is the input dimension and m is the output dimension. β_i is the output weight and $G(a_i, b_i, x_i) = g(a_i \cdot x_i + b_i)$ is the output of the ith hidden neuron where a_i is the input weight vector, b_i is the bias and $g(x)$ is the activation function. Generally, a and b are initialized randomly, the activation function $g(x)$ can be sigmoid function. The training set $\{(x_i, l_i)\,|i = 1, 2, \ldots, N\}$ includes N input vectors and output vectors. Training the model is equivalent to solving a least-squares solution of the linear system: $H\beta = T$, where

$$\boldsymbol{H} = \begin{bmatrix} G(a_1, b_1, x_1) & \ldots & G(a_L, b_L, x_1) \\ \vdots & \ddots & \vdots \\ G(a_1, b_1, x_N) & \ldots & G(a_L, b_L, x_N) \end{bmatrix}_{N\times L} \quad (2)$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L\times m}, \ \boldsymbol{T} = \begin{bmatrix} l_1^T \\ \vdots \\ l_N^T \end{bmatrix}_{N\times m} \quad (3)$$

By minimizing the error between the real and expected output: $\boldsymbol{\beta} = \mathrm{argmin} H\beta - T$, the solution is:

$$\boldsymbol{\beta} = \boldsymbol{H}^{\dagger}\boldsymbol{T} \quad (4)$$

$\boldsymbol{H}^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix.

Based on the spectral graph theory, we construct a neighbor graph G using the KNN (K-Nearest Neighbor) approach. The weight W_{ij} indicates the similarity between vertices v_i and v_j in G, which will be further analyzed in next section. We set Eq.(5) as the objective function to preserve the neighbor relations (*i.e.*, smoothen the manifold), where λ is a relative weight. v_i and v_j are the observed function values on vertex i and j respectively.

$$\min \lambda \sum_{i,j=1}^{l+u} \|v_i - v_j\|_2^2 W_{ij} \tag{5}$$

The objective function can be rewritten as:

$$\min \lambda v^{\mathrm{T}} L v \tag{6}$$

where L is the Laplacian matrix which can be computed as $L = D - W$, and D is the diagonal matrix given by Eq.(7):

$$D_{ii} = \sum_{j=1}^{l+u} W_{ij} \tag{7}$$

2) The construction of a composite neighbor graph

The conventional way of constructing G in SELM is to connect every two-neighbor RSS data items. This does not consider the location information of labeled data. The physical location information can help to recognize the prior information for semi-supervised manifold alignment, which is important for constructing a neighbor graph. The corresponding weights are as follows:

$$W_r(i,j) = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\theta_r}}, x_i \text{ and } x_j \text{ are connected} \\ 0, \qquad \text{otherwise} \end{cases} \tag{8}$$

$$W_{lol}^{l}(i,j) = \begin{cases} e^{-\frac{d(l_i,l_j)^2}{\theta_d}}, l_i \text{ and } l_j \text{ are connected} \\ 0, \qquad \text{otherwise} \end{cases} \tag{9}$$

where $\mathrm{d}(l_i, l_j)$ is the Euclidean distance between location coordinate l_i and l_j, θ_r and θ_d are the kernel parameters which will be determined by the experimental test. The weight matrix W_r is constructed by all $(l+u)$ RSS data and W_{lol}^{l} is constructed by l labeled locations. Then we represent W_r as Eq.(10) and the weight matrix for labeled and unlabeled RSS data as Eq.(11):

$$W_r = \begin{bmatrix} W_r^l & W_r^{lu} \\ W_r^{ul} & W_r^u \end{bmatrix}_{(l+u)\times(l+u)} \tag{10}$$

$$\begin{cases} W^l = \alpha W_r^l + (1-\alpha) W_{lol}^l \\ W^u = W_r^u \end{cases} \tag{11}$$

where $\alpha \in [0,1]$ is a trade-off coefficient. Then we construct the final composite Laplacian matrix L_c:

$$L_c = \begin{bmatrix} L^l & 0 \\ 0 & L^u \end{bmatrix} + \mu L_r^{lu} \tag{12}$$

$$L^l = D^l - W^l,\ L^u = D^u - W^u \tag{13}$$

where μ is the parameter for Laplacian operator, W^l is the weight of l labeled RSS data, W^u is the weight of u unlabeled RSS data respectively, D^l and D^u are the diagonal matrixes computed like Eq.(7), and L_r^{lu} is constructed between l labeled RSS data and u unlabeled RSS data, see Ref.[5] for details.

3. CSELM-QE model

1) CSELM: Composite neighbor graph with SELM

We combine SELM with the above-described composite neighbor graph to optimize localization by minimizing the sum of the square loss function and the smoothness penalty as Eq.(14):

$$\arg\min_{q} \frac{1}{2}\left\{v - T^2 + \lambda v^T L_c v\right\} \tag{14}$$

plug $v = H\beta$ caused by ELM from Eq.(4), we get:

$$\underset{\beta}{\operatorname{argmin}} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\left\{JH\beta - T^2 + \lambda(H\beta)^T L_c H\beta\right\} \tag{15}$$

For the convenience of calculation, J is imported as a diagonal matrix, where the first l diagonal entries are 1 and others are 0. Similarly, the first l elements of T are the true coordinates and the rest u elements are 0. By setting the derivative of objective function to zero, we have $\beta = \left(J + \lambda L_c^T\right)H)^{-1} JT$ and then the outputs of unlabeled RSS data can be estimated.

In the above algorithm, the unlabeled RSS data are collected by walking along the corridors and can be easily affected by environmental changes. Therefore, not all the unlabeled data are trusted. We use the quality estimation method to choose the credible unlabeled RSS data and filter the distrusted data to improve the model accuracy. We first use the Affinity propagation (AP) clustering algorithm[23] to put all the unlabeled RSS data into clusters. For each unlabeled RSS data item, we calculate the quality and filter the data whose quality value is lower than a predetermined threshold in every cluster.

2) CSELM with quality estimation

We use the outlier detection method to filter the unlabeled RSS data that are far away from the cluster centroids. We define the distance threshold r to determine if a data item is the neighbor to another. For each data $x_{j,k} = \{rss_{i1}, rss_{i2}, \ldots, rss_{iM}\}$ in a cluster $X_k = \{x_{1,k}, x_{2,k}, \ldots, x_{n,k}\}$, we calculate the number of its neighbors within the distance threshold. If most of the data items are far away from x_k, then x_k is regarded as an outlier. The equation is as follows:

$$\frac{\{x_{j,k}|dist\left(x_{i,k},x_{j,k}\right)\leq r\}}{|X_k|}\leq\gamma \tag{16}$$

where *dist* is the Euclidean distance between two data items, threshold γ is a fraction. By calculating the number of neighbors, we can delete the outliers for quality estimation. The outliers will be saved into the abnormal data set N_k^A and others into N_k^N.

For all data in N_k^N, we need to calculate the data qualities $Q_k=\{q_{1,k},q_{2,k},\ldots,q_{m,k}\}$ in the kth cluster. We define the centroid of a cluster as ω_k that minimizes the sum of the weighted squared distance between ω_k and other data. This is shown in Eq.(17):

$$\omega_k=\underset{\omega_k}{\operatorname{argmin}}\sum_{i=1}^{m}\left\{dist^2\left(\omega_k,x_{i,k}\right)\times q_{i,k}\right\} \tag{17}$$

The quality is determined on the basis of the deviation $d_{i,k}$ from the cluster centroid, data with higher quality is closer to the centroid. We then set the quality threshold ξ to filter the unlabeled RSS data with lower quality.

$$d_{i,k}=\ dist^2\left(w_k,x_{i,k}\right) \tag{18}$$

The distance threshold r, threshold γ and quality threshold ξ are set by single variable method according to experimental result in Section IV. The pseudo code is shown in Algorithm 1.

The use of ε is to make sure that the equation makes sense when $d_{i,k}=0$. By selection, the unlabeled RSS data with higher quality will be saved as the final selected data set X_c for the CSELM algorithm. Our final radio map consists of limited labeled RSS data and high quality unlabeled RSS data, which compensate for the sparsity of raw radio map and enhance its robustness.

IV. Experiment and Analysis

1. Experimental setup

The experiment is implemented on the 3*rd* floor of No.13 teaching building in Shandong University of Science and Technology, the area of corridors is about 60m by 30m. We developed an application for a mobile device (Android 4.4) to collect the RSS data. For the training data set, we collected 20 samples at each location and there are 2 seconds of time interval between two consecutive RSS measurements. This method can reduce the i of signal fluctuation. These data are used as the labeled data, and the unlabeled data are collected by walking along a predefined path. There are in total 30 APs selected to cover the whole positioning area and a Samsung Galaxy S4 smart phone is selected as the receiver to collect the RSS data. In the experiments, although we do not need all the training data in the actual implementation, we collected all labeled RSS data at 100 reference points and used different percentages of them as a training data set to verify the performance of our proposed algorithm. And we collected another 150 labeled RSS data as test points and 1200 unlabeled data by walking along a predefined path in the corridor.

Algorithm 1 Quality estimation of unlabeled RSS data

```
Input: Clusters of unlabeled RSS data set $X_k = \{x_{1,k},x_{2,k},\ldots,x_{n,k},k=1,2,\ldots,K\}$;
Output: Selected data set $X_c$;
//Outlier Detection
Initialize $N_k^N \leftarrow \Phi, N_k^A \leftarrow \Phi$
for $i \leftarrow 1$ to n do
  count ← 0
  for $j \leftarrow 1$ to n and $j \neq i$ do
    if $dist\,(x_{i,k},x_{j,k}) \leq r$ then
      $count \leftarrow count + 1$
  end
  if $count \geq \gamma n$ then $N_k^N \leftarrow N_k^N \cup \{i\}$
   else $N_k^A \leftarrow N_k^A \cup \{i\}$
end
//Quality estimation process
$q_{i,k} \leftarrow \frac{1}{m}, i \in N_k^N$
while not converged do
   $\omega_k = \arg\min_{\omega_k} \sum_{i=1}^{m} \left\{dist^2\left(\omega_k,x_{i,k}\right) \times q_{i,k}\right\}$
   for each $i \in N_k^N$
     $d_{i,k} = dist(\omega_k,x_{i,k})^2$
end
$\lambda = \sum_{i=1}^{m} d_{i,k}$
update $q_{i,k} = \frac{\frac{1}{\frac{d_{i,k}}{\lambda}+\varepsilon}}{\sum_{j=1}^{m}\frac{1}{\frac{d_{i,k}}{\lambda}+\varepsilon}}, i \in N_k^N$
end
$X_c \leftarrow \{x_{i,k}\}\ \ if\ \ q_{i,k} > \xi$
```

2. Parameter comparison of neighbor weight

In order to get the optimal parameters for the construction of the neighbor weight, we conduct the comparative experiments under different parameters. We use the Euclidean distance to evaluate the localization error, the equation is as below:

$$error=\sqrt{\left(p-\hat{p}\right)^2+\left(q-\hat{q}\right)^2} \tag{19}$$

where (p,q) is the real location coordinate and $(\hat{p},\hat{q})$ is the estimated location coordinate. Clearly, the smaller the error is, the higher the localization accuracy is. We compare four parameters including $\mu,\alpha,\theta_r,\theta_d$ for construction of the neighbor weight. The neighbor number k is six. We use 2500 hidden neurons and 50 percent of all the labeled training data set and 1200 unlabeled data set to train the SELM model with a composite neighbor graph matrix, and 150 test points are used for localization to obtain the optimal parameters.

As shown in Fig.2(a), two parameters affect the ratio of weights including the weight for physical locations and the weight for RSS data. By analyzing their impacts, we can choose the most effective parameters to improve localization accuracy. After obtaining the optimal interval of α and μ, we use the single variable method to make α and μ gradient from 0.1 to 1 and 1 to 10 respectively. When $\alpha = 0.8$ and $\mu = 10$, we obtain the lowest localization error.

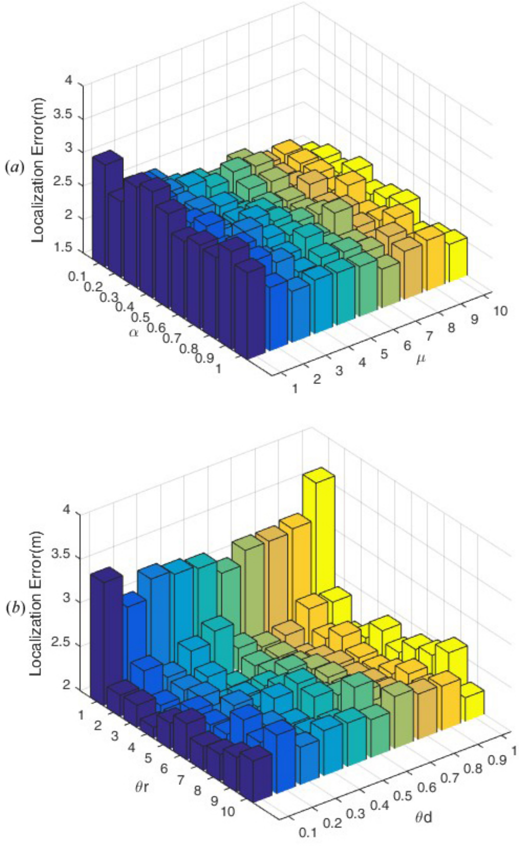


Fig. 2. Localization error under different parameters. (a) Localization error vs. parameters α and μ; (b) Localization error vs. parameters θ_r and θ_d

In addition, for the selection of kernel parameters θ_r and θ_d, we conduct several experiments and obtain the localization error as shown in Fig.2(b). θ_r and θ_d adopt the same method as α and μ to obtain the optimal value. The localization error decreases and then fluctuates with the increases of value θ_r, and the lowest error is obtained when $\theta_r = 6$. And we set $\theta_d = 0.5$ to acquire higher localization accuracy according to the final experimental results.

3. Unlabeled RSS data quality analysis

1) The parameters of RSS data quality estimation

Because we collect the unlabeled RSS data by walking along the corridors, the collected RSS data are not all effective for localization due to signal fluctuation. We conduct 5 random experiments and get the final average result to reveal the effect of different numbers of unlabeled data. We use 10 percent of the labeled reference points as the training set and different numbers (0, 50, 100,...,1200) of unlabeled RSS data to train the CSELM model to obtain the locations of 100 test points. Fig.3 presents the average results of the localization error with respect to the number of unlabeled data. As unlabeled RSS data increases, the localization error shows irregular changes. This result reveals that not all the unlabeled data is effective for localization. Therefore, it is necessary to evaluate the quality of all unlabeled RSS data and select the high quality data to improve localization accuracy.

In the process of quality estimation, we first use the Affinity Propagation clustering algorithm to cluster all the unlabeled RSS data according to RSS similarity. This is computed from the Euclidean distance. For each cluster, we evaluate the data quality using our Quality Estimation method and obtain the final preserved unlabeled data set for localization. In the quality estimation, there are two parameters affecting the selection of unlabeled data. We use 50 percent of the labeled reference points as the training set and all unlabeled RSS data to obtain the optimal parameters for our proposed CSELM-QE algorithm. We conduct the comparative experiments between γ and r which affect the neighbor selection and neighbor number. After obtaining the optimal interval of γ and r we use the single variable method to make γ and r gradient from 0.1 to 0.9 and 2 to 20 respectively. Finally, the lowest localization error is obtained when $\gamma = 0.8$ and $r = 20$.

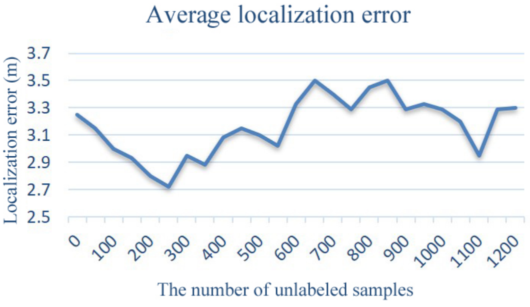


Fig. 3. The localization error with respect to the number of unlabeled RSS data

2) Experimental results of quality estimation

In order to verify the effectiveness of the Quality Estimation method, we use the above-obtained optimal parameters and conduct the comparative experiments of unlabeled RSS data after quality estimation and all unlabeled data. We use different percentages of labeled RSS data to compare the results from selecting the high quality unlabeled data using our proposed method CSELM-QE and all unlabeled data. Fig.4 shows the average localization results against the percentages of labeled RSS data. We can see that our model with the Quality Estimation method performs better no matter how many percentages of labeled RSS data is used compared to the results using all the unlabeled RSS data.

4. Localization performance

To verify the localization performance of using the proposed radio map and other existing methods constructed by SELM[12], GrassMA[5], Kriging[7], LGP[8] and fingerprint-based method, we use 10 percent of labeled RSS data and all unlabeled data to do the localization. The results are shown in Fig.5. In Fig.5(a),it is clear that our proposed CSELM algorithm outperforms the others. Besides, we add the Quality Estimation to SELM and GrassMA, which both use unlabeled RSS data. In Fig.5(b), we can see that the SELM-QE and GrassMA-QE outperform the original SELM and GrassMA methods respectively. The improvement for GrassMA is not obvious because its model is not very dependent on unlabeled RSS data, but SELM improves remarkably. The main probability of localization error is shown in Table 1.

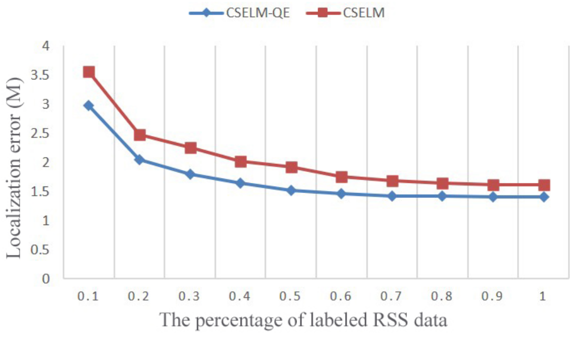


Fig. 4. Average localization error against the percentage of labeled RSS data

Table 1. Comparison of different algorithms under 10 RPs

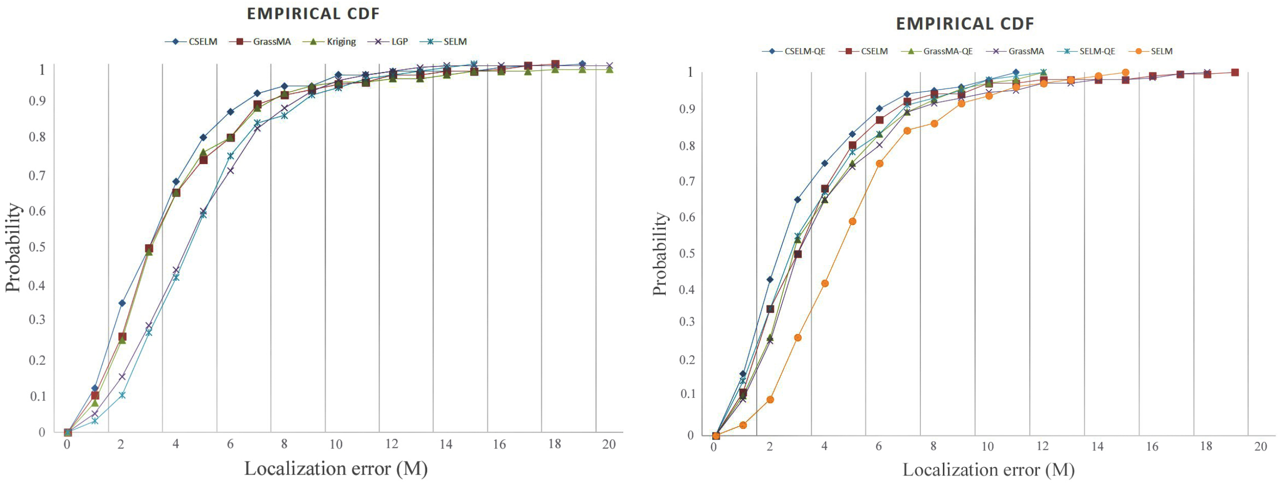
Algorithm	Cumulative probability(Error < 2m)	Average error(m)	CEP(m)	CE95(m)
CSELM	36.8%	3.23	2.80	8.03
SELM	11.2%	4.72	4.29	10.14
GrassMA	25.8%	3.87	2.85	10.27
Kriging	26.2%	4.39	3.09	9.18
LGP	15.3%	4.69	4.51	9.22
Fingerprint-based method	0%	13.97	12.02	22.96
CSELM-QE	42.6%	2.87	2.20	7.51
SELM-QE	35.8%	3.42	2.79	8.91
GrassMA-QE	28.0%	3.63	3.02	8.43

Table 2. Average localization error under different percentages of labeled data

Percentage	CSELM-QE	CSELM	SELM	GrassMA	Kriging	LGP	Fingerprint-based
10%	2.97	3.56	4.57	4.43	4.26	5.16	15.67
20%	2.05	2.47	3.41	3.27	2.47	3.66	9.72
30%	1.79	2.25	2.86	2.91	2.15	2.25	5.29
40%	1.64	2.01	2.69	2.07	1.92	2.04	3.61
50%	1.52	1.92	2.41	1.82	1.81	1.67	2.37
60%	1.46	1.75	2.57	1.61	1.68	1.53	2.34
70%	1.42	1.68	2.46	1.53	1.67	1.52	2.11
80%	1.42	1.64	2.26	1.56	1.62	1.50	1.96
90%	1.41	1.62	2.29	1.52	1.51	1.50	1.82
100%	1.41	1.62	2.25	1.51	1.49	1.49	1.81

The Circular error probable (CEP) indication is defined as the smallest error radius of circle centred at the origin that encloses 50% of test data set. And Circular error 95% (CE95) is similar to CEP that encloses 95% of test data set. First, our CSELM method improves localization accuracy than the SELM algorithm about 31.5%. Second, we combine the Quality Estimation method with CSELM, which improves localization accuracy about 39.2% compared to SELM. This result reveals that the Quality Estimation method is effective for improving localization accuracy. Compared with the basic fingerprint-based method, our algorithm improves about 79.5% without increasing the calibration effort for labeled RSS data. For Average error, CEP and CE95, our CSELM-QE outperforms others.

In addition, we do five random experiments and obtain the average localization error to demonstrate the localization performance based on different percentages of labeled RSS data. We choose labeled RSS data in different percentages randomly and use them for different methods. The localization error gradually decreases as the labeled RSS data increases. As shown in Table 2, our proposed method performs better no matter what percentage of labeled data is used.



(*a*) (*b*)

Fig. 5. The CDF of localization error using different methods; (*a*) Localization error between different methods; (*b*) Localization error after using Quality Estimation.

V. Conclusions

This paper investigated a new method to reduce the effort in radio map construction without sacrificing localization accuracy. By incorporating the physical relations of labeled RSS data, the constructed neighbor graphs include more information for semi-supervised learning. The effects of noise measurements caused by the unlabeled RSS data are reduced by adding a quality estimation method. These high quality data help in improving localization accuracy. Experimental results verify that combining composite neighbor graph and quality estimation can bring good performance on localization accuracy and robustness especially when the labeled data is sparse. In addition, our Quality Estimation method can be applied to other methods which need to retain high quality unlabeled RSS data to improve model accuracy. In the future, we intend to apply our method to the unsupervised approach to further reduce the calibration effort while keeping high localization accuracy.

References

[1] K. Dong, Z. Ling and X. Xia, "Dealing with Insufficient Location Fingerprints in Wi-Fi Based Indoor Location Fingerprinting", *Wireless Communications and Mobile Computing*, pp.1–11, 2017.

[2] P. Bahl and V.N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system", *Proc. of IEEE INFOCOM 2000 Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, Tel Aviv, Israel, pp.775–784, 2000.

[3] B. Hu, H. Peng and Z. Sun, "LANDMARC localization algorithm based on weight optimization", *Chinese Journal of Electronics*, Vol.27, No.6, pp.1291–1296, 2018.

[4] C. Zhu, J. Jia, *et al.*, "Indoor positioning algorithm based on fusion of map information and WiFi landmark", Journal of Shandong University of Science and Technology(Social Sciences), Vol.39, No.1, pp.91–99, 2020.

[5] M. Zhou, Y. Tang, W. Nie, *et al.*, "GrassMA: Graph-based semi-supervised manifold alignment for Indoor WLAN localization", *Sensors*, Vol.17, No.21, pp.7086–7095, 2017.

[6] W. Wu, J. Zhao, C. Zhang, *et al.*, "Improving performance of tensor-based context-aware recommenders using bias tensor factorization with context feature auto-encoding", *Knowledge-Based Systems*, Vol.128, pp.71–77, 2017.

[7] S.S. Jan, S.J. Yeh and Y.W. Liu, "Received signal strength database interpolation by Kriging for a Wi-Fi indoor positioning system", *Sensors*, Vol.15, No.9, pp.21377–21393, 2015.

[8] C. Qiang, L. Qun, S. Zesen, *et al.*, "Scalable indoor localization via Mobile crowdsourcing and gaussian process", *Sensors*, Vol.16, No.3, pp.381–399, 2016.

[9] J. Zhao, X. Gao, X. Wang, *et al.*, "An efficient radio map updating algorithm based on K-Means and gaussian process regression", *Journal of Navigation*, Vol.71, No.5, pp.1055–1068, 2018.

[10] V.K. Jain, S. Tapaswi and A. Shukla, "Location estimation based on semi-supervised locally linear embedding (SSLLE) approach for indoor wireless networks", *Wireless Personal Communications*, Vol.67, No.4, pp.879–893, 2012.

[11] S. Sorour, Y. Lostanlen, S. Valaee, *et al.*, "Joint indoor localization and radio map construction with limited deployment load", *IEEE Transactions on Mobile Computing*, Vol.14, No.5, pp.1031–1043, 2015.

[12] J. Liu, Y. Chen, M. Liu, *et al.*, "SELM: Semi-supervised ELM with application in sparse calibrated location estimation", *Neurocomputing*, Vol.74, No.16, pp.2566–2572, 2011.

[13] X. Chai, and Q. Yang, "Reducing the calibration effort for probabilistic indoor location estimation", *IEEE Transactions on Mobile Computing*, Vol.6, No.6, pp.649–662, 2007.

[14] H. Wang, S. Sen. and A. Elgohary, "No need to war-drive: Unsupervised indoor localization", *Proc. of the 10th International Conference on Mobile Systems*, Lake District, pp.197–210, 2012.

[15] S. Jung and D. Han, "Automated construction and maintenance of Wi-Fi radio maps for crowdsourcing-based indoor positioning systems", *IEEE Access*, Vol.6, pp.1764–1777,

2018.

[16] Y. Ye and B. Wang, “RMapCS: Radio map construction from crowdsourced samples for indoor localization”, *IEEE Access*, Vol.6, pp.24224–24238, 2018.

[17] A. Rai, K.K. Chintalapudi, V.N. Padmanabhan, *et al.*, “Zee: Zero-effort crowdsourcing for indoor localization”, *Proc. of the 18th Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, pp.293–304, 2012.

[18] Z. Yang, C. Wu and Y. Liu, “Locating in fingerprint space: Wireless indoor localization with little human intervention”, *Proc. of the 18th Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, pp.269–280, 2012.

[19] C. Wu, Z. Yang and Y. Liu, “Smartphones based crowdsourcing for indoor localization”, *IEEE Transactions on Mobile Computing*, Vol.14, No.2, pp.444–457, 2015.

[20] S. Liu, Z. Zheng, F. Wu, *et al.*, “Context-aware data quality estimation in mobile crowdsensing”, *Proc. of IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, Atlanta, GA, USA, pp.1–9, 2017.

[21] S. Yang, F. Wu, S. Tang, *et al.*, “On designing data quality-aware truth estimation and surplus sharing method for mobile crowdsensing”, *IEEE Journal on Selected Areas in Communications*, Vol.35, No.4, pp.832–847, 2017.

[22] G.B. Huang, Q.Y. Zhu and C.K. Siew, “Extreme learning machine: A new learning scheme of feedforward neural networks”, *Proc. of IEEE International Joint Conference on Neural Networks*, Budapest, Hungary, pp.985–990, 2004.

[23] B.J Frey and D. Dueck, “Clustering by passing messages between data points”, *Science*, Vol.315, No.5814, pp.972–976, 2007.

ZHAO Jianli received the Ph.D. degree in 2006 from Northeastern University, China. In 2019, He served as a professor in the College of Computer Science and Engineering, Shandong University of Science and Technology. His major research field is pervasive computing and personalized recommendation.
(Email: jlzhao@sdust.edu.cn)

WANG Wei received the B.E. degree in College of Computer Science and Engineering from Shandong University of Science and Technology, Qingdao, China, in 2018. He is studying for the M.E. degree.

SUN Qiuxia (corresponding author) received the Ph.D. degree in 2011 from Qingdao University, China. In 2014, she served as associate professor in College of Mathematics and System Science, Shandong University of Science and Technology. Her major research is big data analysis and complex system Modeling.
(Email: qiuxiasun@sdust.edu.cn)